Day 8: Text Analysis

ME414: Introduction to Data Science and Big Data Analytics

LSE Methods Summer Programme

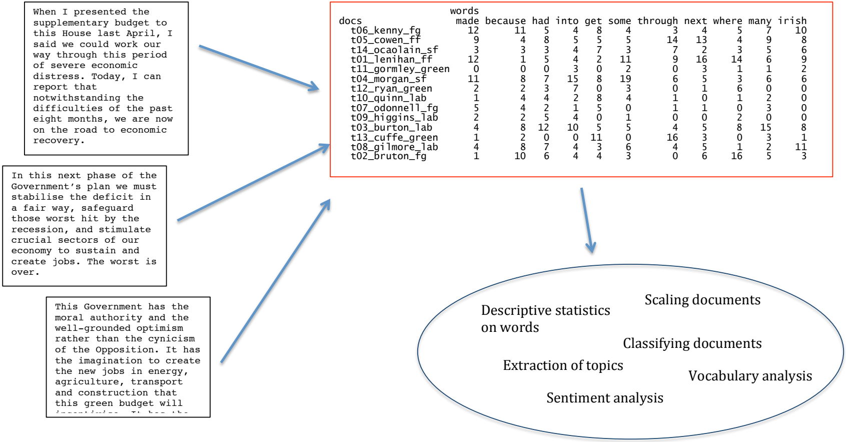
23 August 2017

Day 8 Outline

- Key features of QTA
  - Quantitative text analysis workflow
  - Key basic concepts
- Documents and features
  - Strategies for selecting documents
  - Defining features
  - Parts of speech
  - Filtering features
  - “stopwords”
- Descriptive text analysis
  - Key words in context
  - Descriptive text statistics
  - Lexical diversity
- Content analysis
- Dictionary analysis

Key features of Quantitative Text Analysis

Basic QTA Process: Texts → Feature matrix → Analysis



```
When I presented the
supplementary budget to
this House last April, I
said we could work our
way through this period
of severe economic
distress. Today, I can
report that
notwithstanding the
difficulties of the past
eight months, we are now
on the road to economic
recovery.
```

```
In this next phase of the
Government's plan we must
stabilise the deficit in
a fair way, safeguard
those worst hit by the
recession, and stimulate
crucial sectors of our
economy to sustain and
create jobs. The worst is
over.
```

```
This Government has the
moral authority and the
well-grounded optimism
rather than the cynicism
of the Opposition. It has
the imagination to create
the new jobs in energy,
agriculture, transport
and construction that
this green budget will
```

```
                   words
docs                made because had into get some through next where many irish
  t06_kenny_fg        12      11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff         9       4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf      3       3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff      12       1   5    4   2   11       9   16    14    6     9
  t11_gormley_green    0       0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf       11       8   7   15   8   19       6    5     3    6     6
  t12_ryan_green       2       2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab        1       4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg      5       4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab      2       2   5    4   0    1       0    0     2    0     0
  t03_burton_lab       4       8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green      1       2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab      4       8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg        1      10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

What role for “qualitative” analysis in QTA?

- Ultimately all reading of texts is qualitative, even when we count elements of the text or convert them into numbers
- QTA may involve human judgment in the construction of the feature-document matrix
- QTA may involve human judgment in the interpretation of the output of statistical models

What role for “qualitative” analysis in QTA?

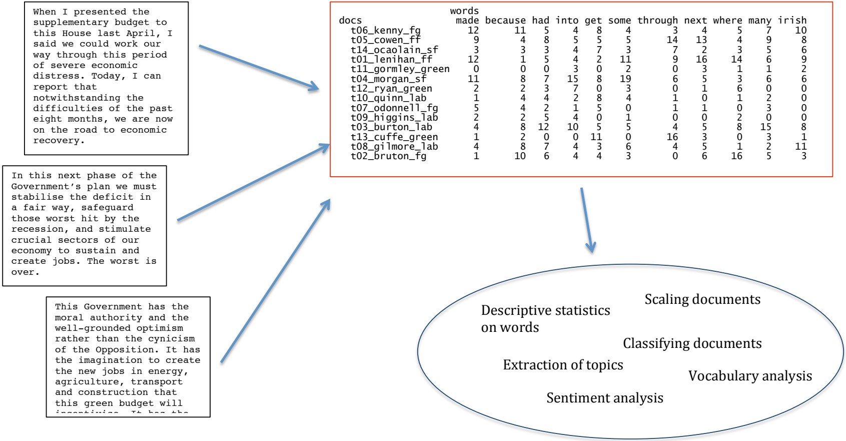
- Ultimately all reading of texts is qualitative, even when we count elements of the text or convert them into numbers
- QTA may involve human judgment in the construction of the feature-document matrix
- QTA may involve human judgment in the interpretation of the output of statistical models
- But quantitative text analysis differs from more qualitiative approaches in that it:
  - Involves large-scale analysis of many texts, rather than close readings of few texts
  - Requires no *interpretation* of texts
- Uses a variety of statistical techniques to extract information from the document-feature matrix

Key feature of quantitative text analysis

- Conversion of textual features into a quantitative matrix.
- A quantitative or statistical procedure to extract information from the quantitative matrix
- Summary and interpretation of the quantitative results

3 guiding priciples for QTA

- All quantitative models for text are wrong, but some are useful
- Quatitative models for text augment, but do not replace, humans
- Validation is key
    - On which note. . . `https://jblumenau.shinyapps.io/topicapp/`



When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will

```
                     words
docs                  made because had into get some through next where many irish
  t06_kenny_fg          12      11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff           9       4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf        3       3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff        12       1   5    4   2   11       9   16    14    6     9
  t11_gormley_green      0       0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf         11       8   7   15   8   19       6    5     3    6     6
  t12_ryan_green         2       2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab          1       4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg        5       4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab        2       2   5    4   0    1       0    0     2    0     0
  t03_burton_lab         4       8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green        1       2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab        4       8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg          1      10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

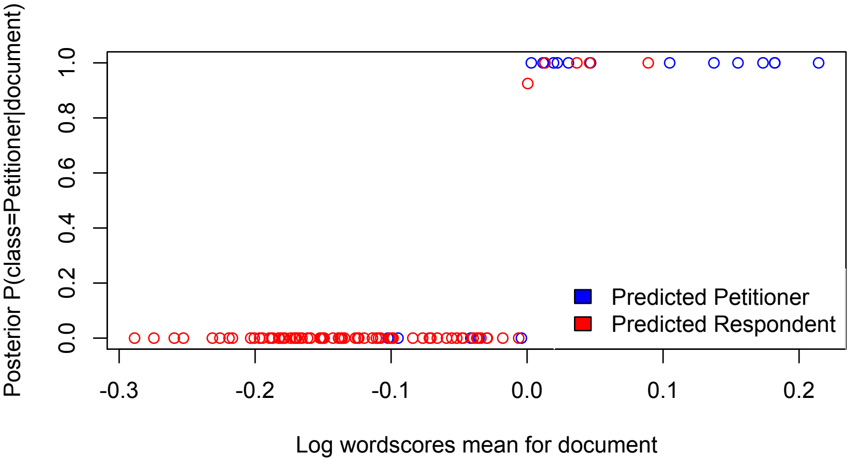
Classifying documents

Extraction of topics

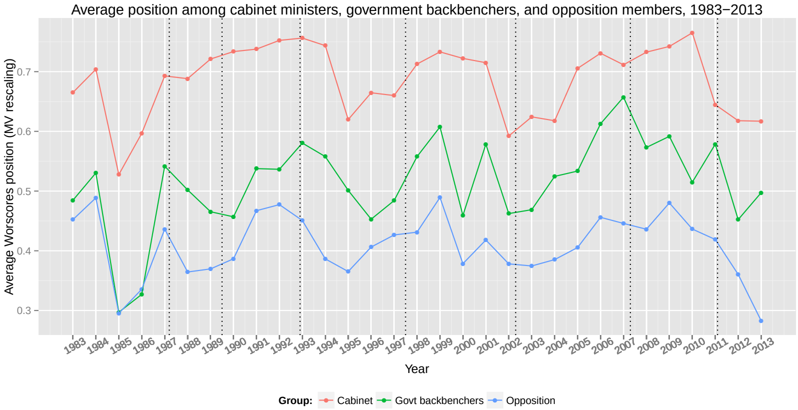
Vocabulary analysis

Sentiment analysis

Example: Document classification using the “Naive Bayes” classifier

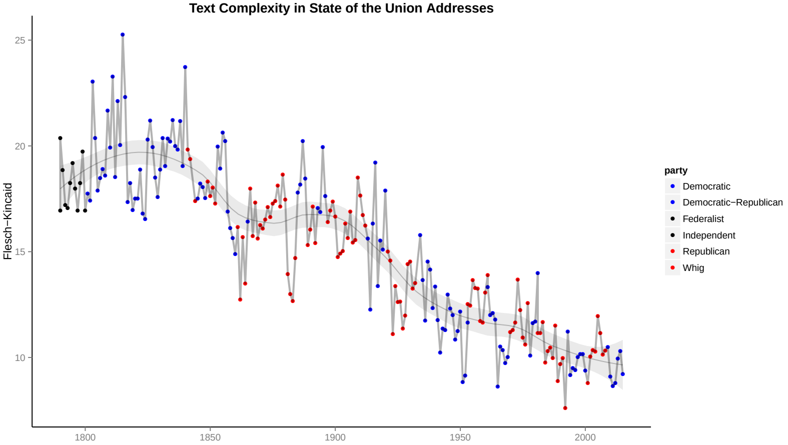


Government v. Opposition in yearly budget debates

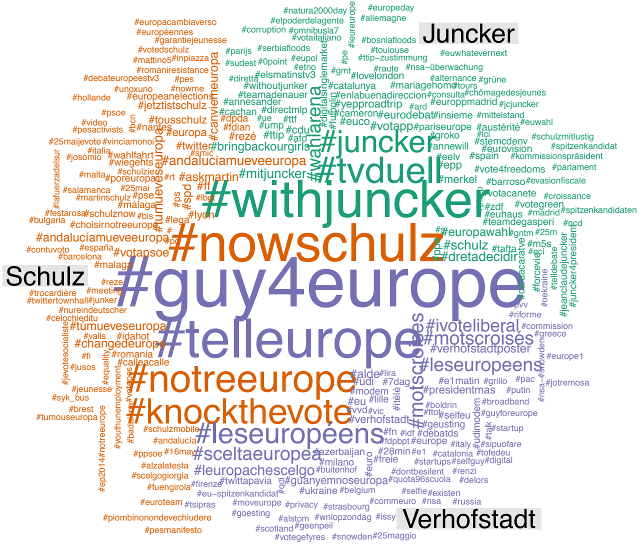

Average position among cabinet ministers, government backbenchers, and opposition members, 1983−2013

(from Herzog and Benoit EPSA 2013)

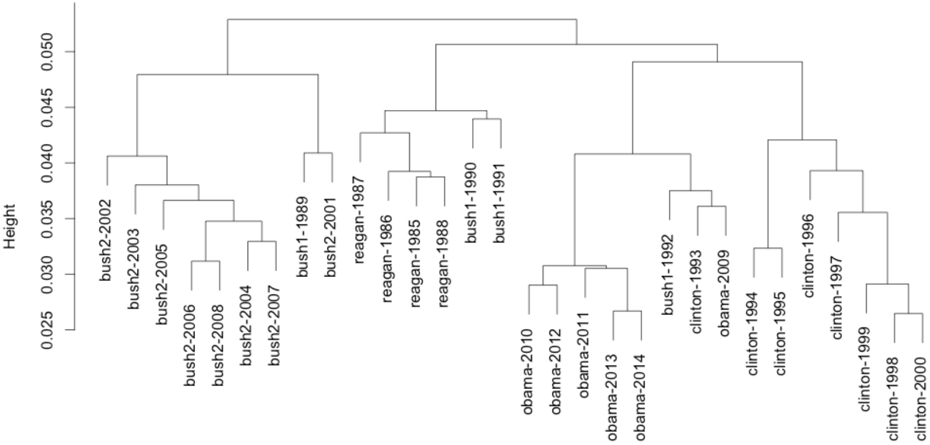
Reading level of US State-of-the-Union addresses over time



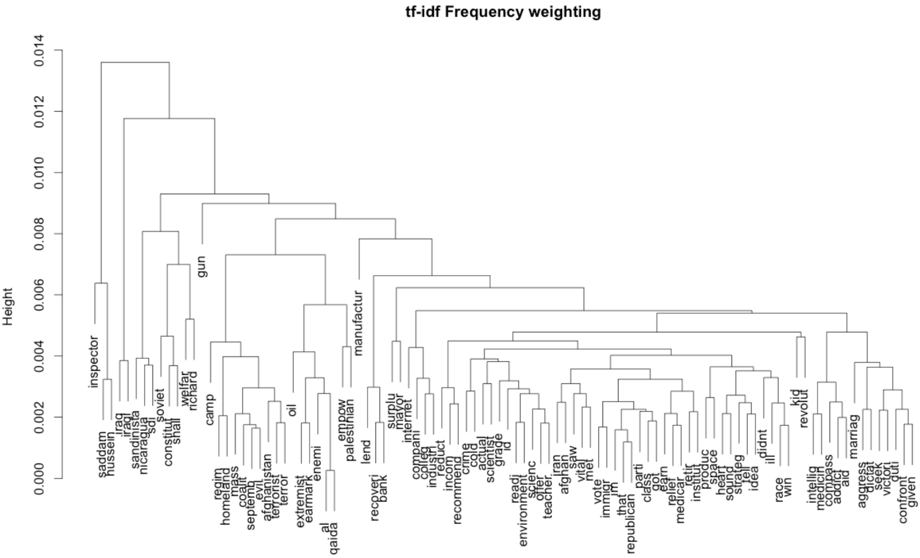
Wordcloud of Tweets from 2014 EP campaign, by list-leading candidate



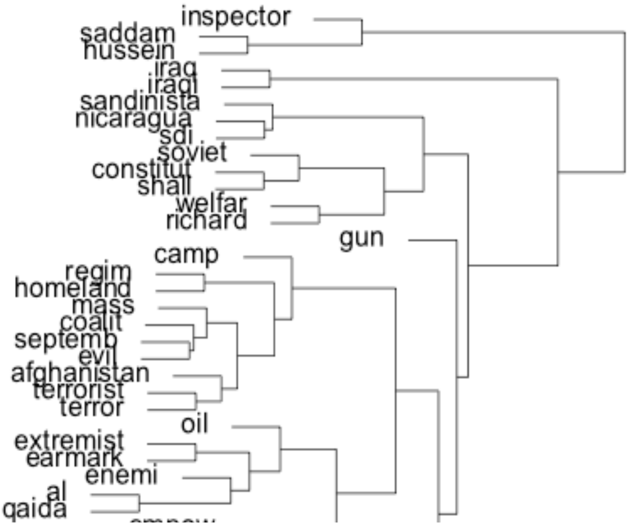
Hierachical clustering: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



Dendrogram: Presidential State of the Union addresses



This requires assumptions

- That texts represent an observable implication of some underlying characteristic of interest (usually an attribute of the author)
- That texts can be represented through extracting their *features*
    - most common is the bag of words assumption
    - disregard grammar, disregard word order, just pay attention to word frequencies
    - many other possible definitions of “features”
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Bag of words assumption

- Consider two sentences:
  1. Time flies like an arrow.
  2. Fruit flies like a banana.
- Convert these into a bag-of-words feature matrix:

Bag of words assumption

- Consider two sentences:
  1. Time flies like an arrow.
  2. Fruit flies like a banana.
- Convert these into a bag-of-words feature matrix:

	time	flies	fruit	like	an	a	banana	arrow
Sentence 1	1	1	0	1	1	0	0	1
Sentence 2	0	1	1	1	0	1	1	0

Bag of words assumption

- Consider two sentences:
  1. Time flies like an arrow.
  2. Fruit flies like a banana.
- Convert these into a bag-of-words feature matrix:

	time	flies	fruit	like	an	a	banana	arrow
Sentence 1	1	1	0	1	1	0	0	1
Sentence 2	0	1	1	1	0	1	1	0

- The dependency structure between words in each sentence is lost
- The word “flies” has a different meaning in the two sentences (metaphorical versus literal), but both sentences score a 1 here
- The ‘joke’ is no longer funny

Key feature of quantitative text analysis

1. Selecting texts: Defining the *corpus*
2. Conversion of texts into a common electronic format
3. Defining documents: deciding what will be the doumentary unit of analysis

Key feature of quantitative text analysis (cont.)

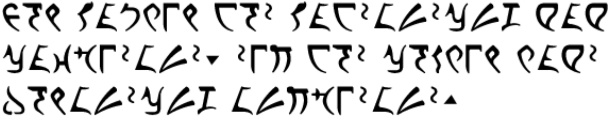
4. Defining features. These can take a variety of forms, including tokens, equivalence classes of tokens (dictionaries), selected phrases, human-coded segments (of possibily variable length), linguistic features, and more.
5. Conversion of textual features into a quantitative matrix
6. A quantitative or statistical procedure to extract information from the quantitative matrix
7. Summary and interpretation of the quantitative results

Extreme forms of QTA

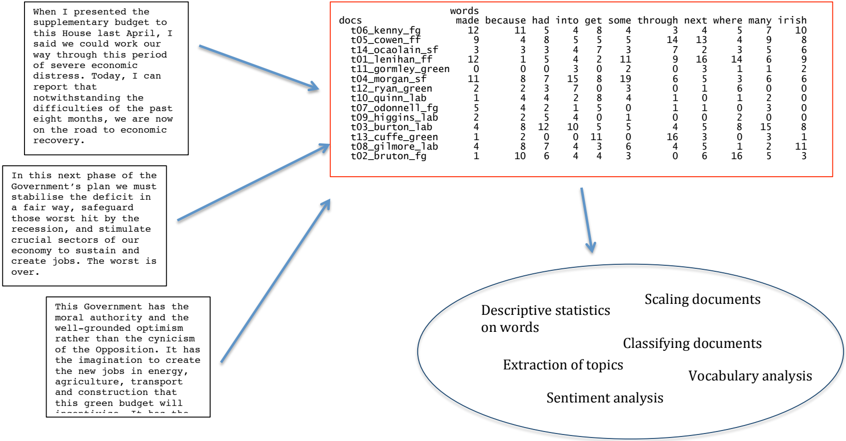
- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Methods can “discover” topics with little human supervision
- Language-blind: can scaling anything that occurs with regular patterns (even without knowing what these mean)

Extreme forms of QTA

- Fully automated technique with minimal human intervention or judgment calls – only with regard to reference text selection
- Methods can “discover” topics with little human supervision
- Language-blind: can scaling anything that occurs with regular patterns (even without knowing what these mean)
- Could potentially work on texts like this:



(See http://www.kli.org)



When I presented the supplementary budget to this House last April, I said we could work our way through this period of severe economic distress. Today, I can report that notwithstanding the difficulties of the past eight months, we are now on the road to economic recovery.

In this next phase of the Government's plan we must stabilise the deficit in a fair way, safeguard those worst hit by the recession, and stimulate crucial sectors of our economy to sustain and create jobs. The worst is over.

This Government has the moral authority and the well-grounded optimism rather than the cynicism of the Opposition. It has the imagination to create the new jobs in energy, agriculture, transport and construction that this green budget will incentivise. It has the

```
                   words
docs                made because had into get some through next where many irish
  t06_kenny_fg        12      11   5    4   8    4       3    4     5    7    10
  t05_cowen_ff         9       4   8    5   5    5      14   13     4    9     8
  t14_ocaolain_sf      3       3   3    4   7    3       7    2     3    5     6
  t01_lenihan_ff      12       1   5    4   2   11       9   16    14    6     9
  t11_gormley_green    0       0   0    3   0    2       0    3     1    1     2
  t04_morgan_sf       11       8   7   15   8   19       6    5     3    6     6
  t12_ryan_green       2       2   3    7   0    3       0    1     6    0     0
  t10_quinn_lab        1       4   4    2   8    4       1    0     1    2     0
  t07_odonnell_fg      5       4   2    1   5    0       1    1     0    3     0
  t09_higgins_lab      2       2   5    4   0    1       0    0     2    0     0
  t03_burton_lab       4       8  12   10   5    5       4    5     8   15     8
  t13_cuffe_green      1       2   0    0  11    0      16    3     0    3     1
  t08_gilmore_lab      4       8   7    4   3    6       4    5     1    2    11
  t02_bruton_fg        1      10   6    4   4    3       0    6    16    5     3
```

Descriptive statistics on words

Scaling documents

Classifying documents

Extraction of topics

Vocabulary analysis

Sentiment analysis

Some key basic concepts

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

types for our purposes, a unique word

Some key basic concepts

(text) corpus a large and structured set of texts for analysis

types for our purposes, a unique word

tokens any word – so token count is total words

stems words with suffixes removed

lemmas canonical word form (the base form of a word that has the same meaning even when different suffixes (or prefixes) are attached)

keys such as dictionary entries, where the user defines a set of equivalence classes that group different word types

Some more key basic concepts

Some more key basic concepts

"key" words	Words selected because of special attributes, meanings, or rates of occurrence

Some more key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

Some more key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

readability provides estimates of the readability of a text based on word length, syllable length, etc.

Some more key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

readability provides estimates of the readability of a text based on word length, syllable length, etc.

complexity A word is considered “complex” if it contains three syllables or more

Some more key basic concepts

“key” words Words selected because of special attributes, meanings, or rates of occurrence

stop words Words that are designated for exclusion from any analysis of a text

readability provides estimates of the readability of a text based on word length, syllable length, etc.

complexity A word is considered “complex” if it contains three syllables or more

diversity (lexical diversity) A measure of how many types occur per fixed word rate (a normalized vocabulary measure)

Documents and Features

Strategies for selecting units of textual analysis

- Words
- *n*-word sequences
- pages
- paragraphs
- Natural units (a speech, a poem, a manifesto)
- Key: depends on the research design

Defining Features

- words

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g.

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g. *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz*

Defining Features

- words
- word stems or lemmas: this is a form of defining *equivalence classes* for word features
- word segments, especially for languages using compound words, such as German, e.g. *Rindfleischetikettierungsüberwachungsaufgabenübertragungsgesetz* (the law concerning the delegation of duties for the supervision of cattle marking and the labelling of beef)

Defining Features (cont.)

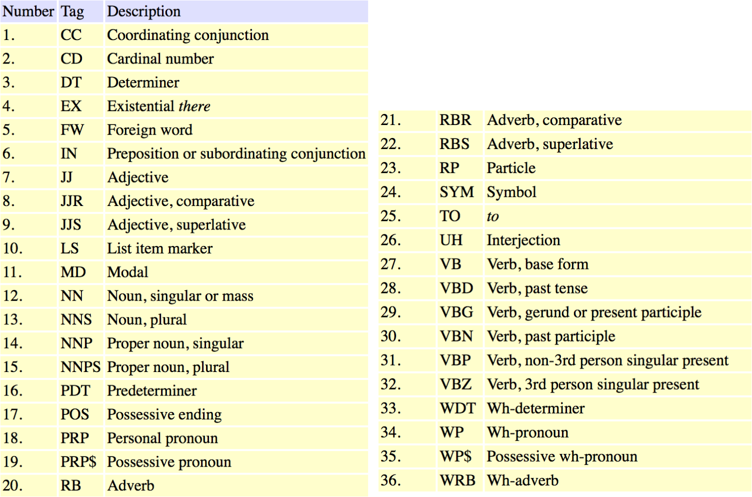
- “word” sequences, especially when inter-word delimiters (usually white space) are not commonly used, as in Chinese

  莎拉波娃现在居住在美国东南部的佛罗里达。今年4月9日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

- n-grams: contiguous sequence of words from document (1-gram, unigram; 2-gram, bigram, etc)
- (if qualitative coding is used) coded or annotated text segments
- linguistic features: parts of speech

Parts of speech

- the Penn “Treebank” is the standard scheme for tagging POS



Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Parts of speech (cont.)

- several open-source projects make it possible to tag POS in text, namely Apache's OpenNLP (and R package openNLP wrapper)

```
> s
Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.
Mr. Vinken is chairman of Elsevier N.V., the Dutch publishing group.
> sprintf("%s/%s", s[a3w], tags)
 [1] "Pierre/NNP"      "Vinken/NNP"      ",/,"             "61/CD"
 [5] "years/NNS"       "old/JJ"          ",/,"             "will/MD"
 [9] "join/VB"         "the/DT"          "board/NN"        "as/IN"
[13] "a/DT"            "nonexecutive/JJ" "director/NN"     "Nov./NNP"
[17] "29/CD"           "./."             "Mr./NNP"         "Vinken/NNP"
[21] "is/VBZ"          "chairman/NN"     "of/IN"           "Elsevier/NNP"
[25] "N.V./NNP"        ",/,"             "the/DT"          "Dutch/JJ"
[29] "publishing/NN"   "group/NN"        "./."
```

Strategies for feature selection

- This can lead to a lot of features!
- An example (small) corpus:
  - 17,129 speeches made in the final month of 2016 in the House of Commons
  - $\approx$ 3 million total words
  - 46998 unique words
  - 468244 unique 1-gram and 2-gram sequences

Strategies for feature selection

- document frequency How many documents in which a term appears
- term frequency How many times does the term appear in the corpus
- deliberate disregard Use of "stop words": words excluded because they represent linguistic connectors of no substantive content
- purposive selection Use of a *dictionary* of words or phrases
- declared equivalency classes Non-exclusive synonyms, what I call a *thesaurus* (lots more on these on Day 4)

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

Common English stop words

a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot, could, dear, did, do, does, either, else, ever, every, for, from, get, got, had, has, have, he, her, hers, him, his, how, however, I, if, in, into, is, it, its, just, least, let, like, likely, may, me, might, most, must, my, neither, no, nor, not, of, off, often, on, only, or, other, our, own, rather, said, say, says, she, should, since, so, some, than, that, the, their, them, then, there, these, they, this, tis, to, too, twas, us, wants, was, we, were, what, when, where, which, while, who, whom, why, will, with, would, yet, you, your

- But no list should be considered universal

A more comprehensive list of stop words

a’s, able, about, above, according, accordingly, across, actually, after, afterwards, again, against, ain’t, all, allow, allows, almost, alone, along, already, also, although, always, am, among, amongst, an, and, another, any, anybody, anyhow, anyone, anything, anyway, anyways, anywhere, apart, appear, appreciate, appropriate, are, aren’t, around, as, aside, ask, asking, associated, at, available, away, awfully, be, became, because, become, becomes, becoming, been, before, beforehand, behind, being, believe, below, beside, besides, best, better, between, beyond, both, brief, but, by, c’mon, c’s, came, can, can’t, cannot, cant, cause, causes, certain, certainly, changes, clearly, co, com, come, comes, concerning, consequently, consider, considering, contain, containing, contains, corresponding, could, couldn’t, course, currently, definitely, described, despite, did, didn’t, different, do, does, doesn’t, doing, don’t, done, down, downwards, during, each, edu, eg, eight, either, else, elsewhere, enough, entirely, especially, et, etc, even, ever, every, everybody, everyone, everything, everywhere, ex, exactly, example, except, far, few, fifth, first, five, followed, following, follows, for, former, formerly, forth, four, from, further, furthermore, get, gets, getting, given, gives, go, goes, going, gone, got, gotten, greetings, had, hadn’t, happens, hardly, has, hasn’t, have, haven’t, having, he, he’s, hello, help, hence, her, here, here’s, hereafter, hereby, herein, hereupon, hers, herself, hi, him, himself, his, hither, hopefully, how, howbeit, however, i’d, i’ll, i’m, i’ve, ie, if, ignored, immediate, in, inasmuch, inc, indeed, indicate, indicated, indicates, inner, insofar, instead, into, inward, is, isn’t, it, it’d, it’ll, it’s, its, itself, just, keep, keeps, kept, know, knows, known, last, lately, later, latter, latterly, least, less, lest, let, let’s, like, liked, likely, little, look, looking, looks, ltd, mainly, many, may, maybe, me, mean, meanwhile, merely, might, more, moreover, most, mostly, much, must, my, myself, name, namely, nd, near, nearly, necessary, need, needs, neither, never, nevertheless, new, next, nine, no, nobody, non, none, noone, nor, normally, not, nothing, novel, now, nowhere, obviously, of, off, often, oh, ok, okay, old, on, once, one, ones, only, onto, or, other, others, otherwise, ought, our, ours, ourselves, out, outside, over, overall, own, particular, particularly, per, perhaps, placed, please, plus, possible, presumably, probably, provides, que, quite, qv, rather, rd, re, really, reasonably, regarding, regardless, regards, relatively, respectively, right, said, same, saw, say, saying, says, second, secondly, see, seeing, seem, seemed, seeming, seems, seen, self, selves, sensible, sent, serious, seriously, seven, several, shall, she, should, shouldn’t, since, six, so, some, somebody, somehow, someone, something, sometime, sometimes, somewhat, somewhere, soon, sorry, specified, specify, specifying, still, sub, such, sup, sure, t’s, take, taken, tell, tends, th, than, thank, thanks, thanx, that, that’s, thats, the, their, theirs, them, themselves, then, thence, there, there’s, thereafter, thereby, therefore, therein, theres, thereupon, these, they, they’d, they’ll, they’re, they’ve, think, third, this, thorough, thoroughly, those, though, three, through, throughout, thru, thus, to, together, too, took, toward, towards, tried, tries, truly, try, trying, twice, two, un, under, unfortunately, unless, unlikely, until, unto, up, upon, us, use, used, useful, uses, using, usually, value, various, very, via, viz, vs, want, wants, was, wasn’t, way, we, we’d, we’ll, we’re, we’ve, welcome, well, went, were, weren’t, what, what’s, whatever, when, whence, whenever, where, where’s, whereafter, whereas, whereby, wherein, whereupon, wherever, whether, which, while, whither, who, who’s, whoever, whole, whom, whose, why, will, willing, wish,

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

Stemming words

Lemmatization refers to the algorithmic process of converting words to their lemma forms.

stemming the process for reducing inflected (or sometimes derived) words to their stem, base or root form. Different from *lemmatization* in that stemmers operate on single words without knowledge of the context.

both convert the morphological variants into stem or root terms

example: `produc` from `production, producer, produce, produces, produced`

Feature selection in practice

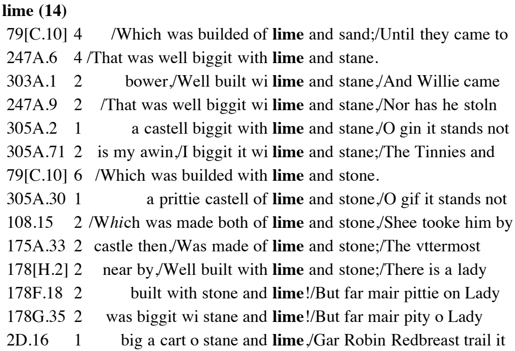
- 17,129 speeches made in the final month of 2016 in the House of Commons
- 46,998 unique words
  - After stopwords: 46,835
  - . . . and stemming: 36,460
  - . . . and removing features that appear fewer than 5 times: 8,823
  - . . . and removing features in fewer than 0.001 documents: 4,068
- Feature selection matters! See Denny and Spirling, 2017
  - Just seven (binary) preprocessing decisions leads to a total of 27 = 128 possible feature matrices
  - These selection decisions can have substantive implications for the inferences we draw from QTA

Descriptive text analysis

Exploring Texts: Key Words in Context

Exploring Texts: Key Words in Context

KWIC *Key words in context* Refers to the most common format for concordance lines. A KWIC index is formed by sorting and aligning the words within an article title to allow each word (except the stop words) in titles to be searchable alphabetically in the index.



lime (14)

79[C.10]	4	/Which was builded of	**lime**	and sand;/Until they came to
247A.6	4	/That was well biggit with	**lime**	and stane.
303A.1	2	bower,/Well built wi	**lime**	and stane,/And Willie came
247A.9	2	/That was well biggit wi	**lime**	and stane,/Nor has he stoln
305A.2	1	a castell biggit with	**lime**	and stane,/O gin it stands not
305A.71	2	is my awin,/I biggit it wi	**lime**	and stane;/The Tinnies and
79[C.10]	6	/Which was builded with	**lime**	and stone.
305A.30	1	a prittie castell of	**lime**	and stone,/O gif it stands not
108.15	2	/W*hi*ch was made both of	**lime**	and stone,/Shee tooke him by
175A.33	2	castle then,/Was made of	**lime**	and stone;/The vttermost
178[H.2]	2	near by,/Well built with	**lime**	and stone;/There is a lady
178F.18	2	built with stone and	**lime**	!/But far mair pittie on Lady
178G.35	2	was biggit wi stane and	**lime**	!/But far mair pity o Lady
2D.16	1	big a cart o stane and	**lime**	,/Gar Robin Redbreast trail it

Another KWIC Example (Seale et al (2006)

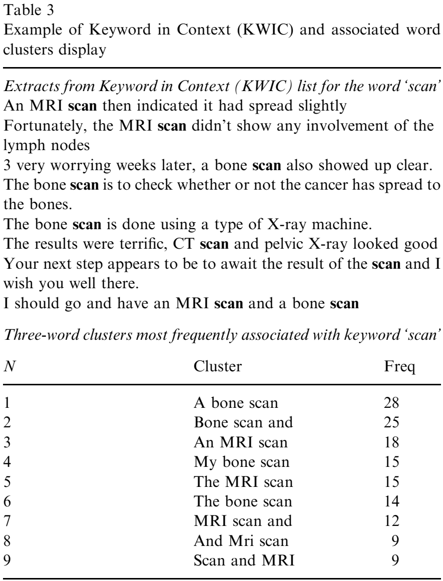


Table 3
Example of Keyword in Context (KWIC) and associated word clusters display

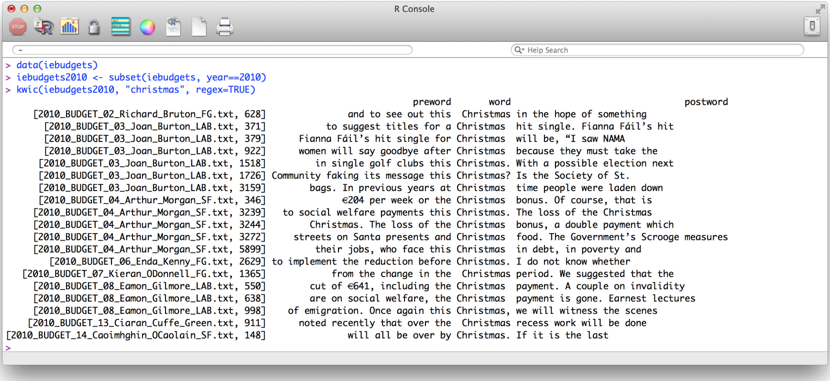
Extracts from Keyword in Context (KWIC) list for the word 'scan'

An MRI **scan** then indicated it had spread slightly
Fortunately, the MRI **scan** didn't show any involvement of the lymph nodes
3 very worrying weeks later, a bone **scan** also showed up clear.
The bone **scan** is to check whether or not the cancer has spread to the bones.
The bone **scan** is done using a type of X-ray machine.
The results were terrific, CT **scan** and pelvic X-ray looked good
Your next step appears to be to await the result of the **scan** and I wish you well there.
I should go and have an MRI **scan** and a bone **scan**

Three-word clusters most frequently associated with keyword 'scan'

N	Cluster	Freq
1	A bone scan	28
2	Bone scan and	25
3	An MRI scan	18
4	My bone scan	15
5	The MRI scan	15
6	The bone scan	14
7	MRI scan and	12
8	And Mri scan	9
9	Scan and MRI	9

Irish Budget Speeches KIWC in quanteda



```
> data(iebudgets)
> iebudgets2010 <- subset(iebudgets, year==2010)
> kwic(iebudgets2010, "christmas", regex=TRUE)
                                                                           preword       word                                 postword
     [2010_BUDGET_02_Richard_Bruton_FG.txt, 628]               and to see out this  Christmas in the hope of something
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 371]           to suggest titles for a Christmas  hit single. Fianna Fáil’s hit
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 379]      Fianna Fáil’s hit single for Christmas  will be, “I saw NAMA
       [2010_BUDGET_03_Joan_Burton_LAB.txt, 922]      women will say goodbye after Christmas  because they must take the
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 1518]         in single golf clubs this Christmas. With a possible election next
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 1726] Community faking its message this Christmas? Is the Society of St.
      [2010_BUDGET_03_Joan_Burton_LAB.txt, 3159]        bags. In previous years at Christmas  time people were laden down
      [2010_BUDGET_04_Arthur_Morgan_SF.txt, 346]              €204 per week or the Christmas  bonus. Of course, that is
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3239]   to social welfare payments this Christmas. The loss of the Christmas
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3244]        Christmas. The loss of the Christmas  bonus, a double payment which
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 3272]     streets on Santa presents and Christmas  food. The Government’s Scrooge measures
     [2010_BUDGET_04_Arthur_Morgan_SF.txt, 5899]         their jobs, who face this Christmas  in debt, in poverty and
        [2010_BUDGET_06_Enda_Kenny_FG.txt, 2629] to implement the reduction before Christmas. I do not know whether
   [2010_BUDGET_07_Kieran_ODonnell_FG.txt, 1365]            from the change in the  Christmas period. We suggested that the
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 550]        cut of €641, including the Christmas  payment. A couple on invalidity
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 638]        are on social welfare, the Christmas  payment is gone. Earnest lectures
     [2010_BUDGET_08_Eamon_Gilmore_LAB.txt, 998]    of emigration. Once again this Christmas, we will witness the scenes
    [2010_BUDGET_13_Ciaran_Cuffe_Green.txt, 911]      noted recently that over the  Christmas recess work will be done
[2010_BUDGET_14_Caoimhghin_OCaolain_SF.txt, 148]               will all be over by Christmas. If it is the last
>
```

Basic descriptive summaries of text

Readability statistics Use a combination of syllables and sentence length to indicate “readability” in terms of complexity

Vocabulary diversity (At its simplest) involves measuring a *type-to-token ratio* (TTR) where unique words are types and the total words are tokens

Word (relative) frequency

Theme (relative) frequency

Length in characters, words, lines, sentences, paragraphs, pages, sections, chapters, etc.

Simple descriptive table about texts: Describe your data!

Speaker	Party	Tokens	Types
Brian Cowen	FF	5,842	1,466
Brian Lenihan	FF	7,737	1,644
Ciaran Cuffe	Green	1,141	421
John Gormley (Edited)	Green	919	361
John Gormley (Full)	Green	2,998	868
Eamon Ryan	Green	1,513	481
Richard Bruton	FG	4,043	947
Enda Kenny	FG	3,863	1,055
Kieran ODonnell	FG	2,054	609
Joan Burton	LAB	5,728	1,471
Eamon Gilmore	LAB	3,780	1,082
Michael Higgins	LAB	1,139	437
Ruairi Quinn	LAB	1,182	413
Arthur Morgan	SF	6,448	1,452
Caoimhghin O'Caolain	SF	3,629	1,035
All Texts		49,019	4,840
Min		919	361
Max		7,737	1,644
Median		3,704	991
Hapaxes with Gormley Edited		67	
Hapaxes with Gormley Full Speech		69	

Lexical Diversity

- Basic measure is the TTR: Type-to-Token ratio

$$TTR = \frac{NumberofTypes(V)}{NumberofTokens(N)} \tag{1}$$

- Problem: This is very sensitive to overall document length, as shorter texts may exhibit fewer word repetitions
- Special problem: length may relate to the introdution of additional subjects, which will also increase richness

Vocabulary diversity and corpus length

- In natural language text, the rate at which new types appear is very high at first, but diminishes with added tokens

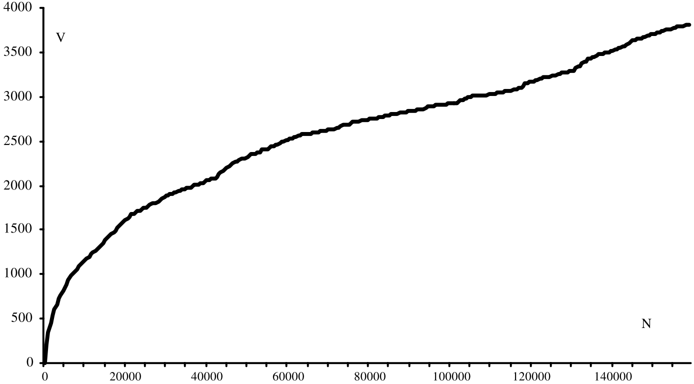


Fig. 1. Chart of vocabulary growth in the tragedies of Racine (chronological order, 500 token intervals).

Vocabulary Diversity Example

- Variations use automated segmentation – here approximately 500 words in a corpus of serialized, concatenated weekly addresses by de Gaulle (from Labbé et. al. 2004)
- While most were written, during the period of December 1965 these were more spontaneous press conferences

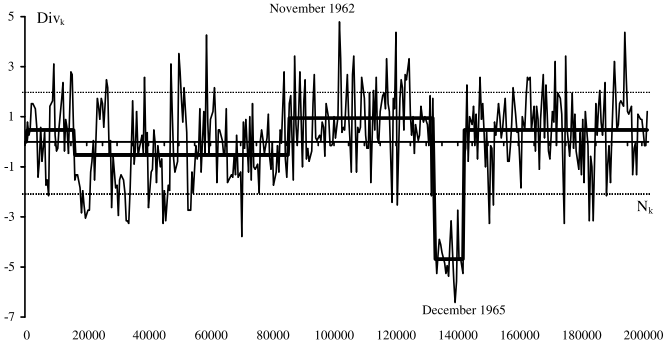
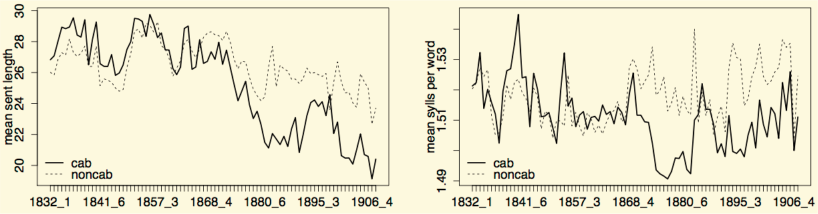


Fig. 8. Evolution of vocabulary diversity in General de Gaulle's broadcast speeches (June 1958–April 1969).

Readability Example

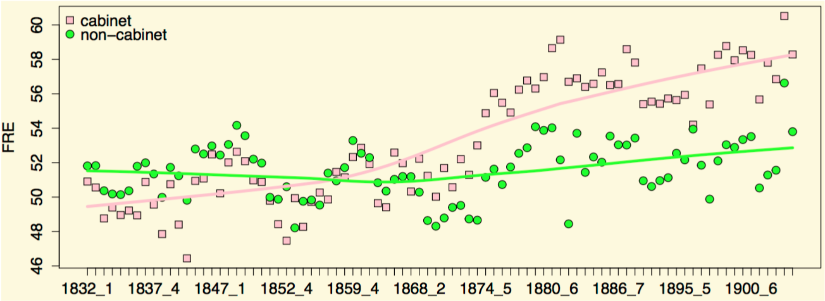
- Most commonly used readability scores focus on a combination of syllables and sentence length
  - Shorter sentences = more readable
  - Fewer syllables = more readable
- Research question: Do Members of Parliament use less complex language when appealing to a more diverse electorate?
- Context: Parliamentary speeches before and after the Great Reform Act (1867)



Readability Example

Flesch score:

$$206.835 - 1.105\left(\frac{\text{total number of words}}{\text{total number of sentences}}\right) - 84.6\left(\frac{\text{total number of syllables}}{\text{total number of words}}\right)$$



Content analysis

Hand-coding: “Classic” content analysis

- Key feature: use of “human” coders to implement a pre-defined coding scheme, by reading and coding texts

- Human decision-making is the central feature of coding decisions, not a computer or other mechanized tool

- Differs from thematic analysis in that the coding scheme is *fixed*

- Alternative 1: (somewhat more automated) is a dictionary approach

- Alternative 2: (entirely “automated”) is inductive scaling of texts from the quantitative matrix

Hand-coding': "Classic" content analysis

- Validity is usually the objective, rather than reliability
- Another motivating factor could be ease of use, or the difficulty of implementing an automated procedure
- May be *computer-assisted*, especially for unitization
- Many common "CATA" or "CACA" tools exist – e.g. QDAMiner

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Summarizing Reducing the coded data to summary quantities of interest.

Components of classical content analysis designs

Unitizing The systematic distinguishing of segments of text that are of interest to the analysis.

Sampling Choice (and justification of the choice) of text units to sample, from population of possible text units.

Coding Classifying each coded unit of text from the sample according to the pre-defined category scheme.

Summarizing Reducing the coded data to summary quantities of interest.

Inference and reporting The final steps wherein the analyzed results are used to generalize about social world, and communicating these results to others.

Dictionary analysis

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings
- Two components:

  key the label for the equivalence class for the concept or canonical term

  values (multiple) terms or patterns that are declared equivalent occurences of the key class
- Frequently involves lemmatization: transformation of all inflected word forms to their "dictionary look-up form" — more powerful than stemming

Rationale for dictionaries

- Rather than count words that occur, pre-define words associated with specific meanings

- Two components:

  key the label for the equivalence class for the concept or canonical term

  values (multiple) terms or patterns that are declared equivalent occurences of the key class

- Frequently involves lemmatization: transformation of all inflected word forms to their “dictionary look-up form” — more powerful than stemming

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- "Qualitiative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment

Bridging qualitative and quantitative text analysis

- A hybrid procedure between qualitative and quantitative classification the fully automated end of the text analysis spectrum
- "Qualitiative" since it involves identification of the concepts and associated keys/categories, and the textual features associated with each key/category
- Dictionary construction involves a lot of contextual interpretation and qualitative judgment
- Perfect reliability because there is no human decision making as part of the text analysis procedure

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented

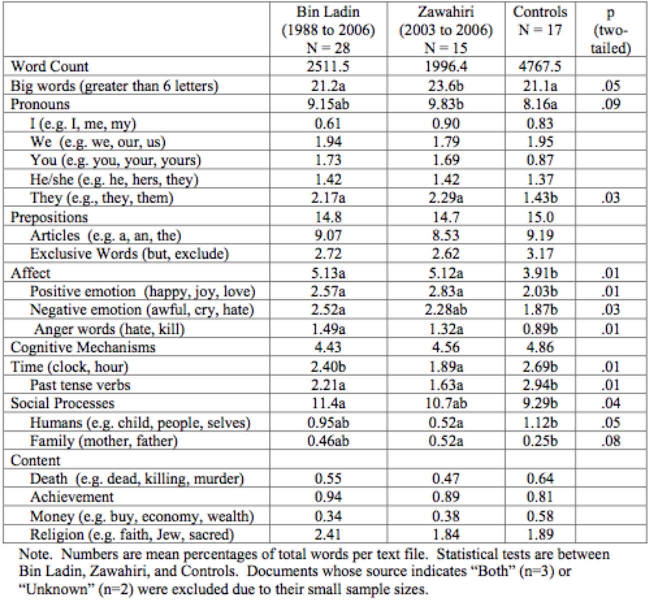
Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory

Linquistic Inquiry and Word Count

- Craeted by Pennebaker et al — see `http://www.liwc.net`
- uses a dictionary to calculate the percentage of words in the text that match each of up to 82 language dimensions
- Consists of about 4,500 words and word stems, each defining one or more word categories or subdictionaries
- For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verb, and past tense verb. So observing the token *cried* causes each of these five subdictionary scale scores to be incremented
- Hierarchical: so "anger" are part of an *emotion* category and a *negative emotion* subcategory
- You can buy it here: `http://www.liwc.net/descriptiontable1.php`

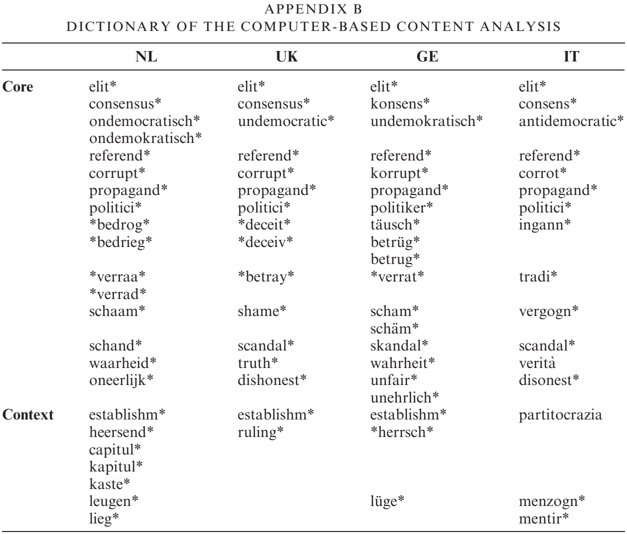
Example: Terrorist speech



	Bin Ladin (1988 to 2006) N = 28	Zawahiri (2003 to 2006) N = 15	Controls N = 17	p (two-tailed)
Word Count	2511.5	1996.4	4767.5	
Big words (greater than 6 letters)	21.2a	23.6b	21.1a	.05
Pronouns	9.15ab	9.83b	8.16a	.09
I (e.g. I, me, my)	0.61	0.90	0.83	
We (e.g. we, our, us)	1.94	1.79	1.95	
You (e.g. you, your, yours)	1.73	1.69	0.87	
He/she (e.g. he, hers, they)	1.42	1.42	1.37	
They (e.g., they, them)	2.17a	2.29a	1.43b	.03
Prepositions	14.8	14.7	15.0	
Articles (e.g. a, an, the)	9.07	8.53	9.19	
Exclusive Words (but, exclude)	2.72	2.62	3.17	
Affect	5.13a	5.12a	3.91b	.01
Positive emotion (happy, joy, love)	2.57a	2.83a	2.03b	.01
Negative emotion (awful, cry, hate)	2.52a	2.28ab	1.87b	.03
Anger words (hate, kill)	1.49a	1.32a	0.89b	.01
Cognitive Mechanisms	4.43	4.56	4.86	
Time (clock, hour)	2.40b	1.89a	2.69b	.01
Past tense verbs	2.21a	1.63a	2.94b	.01
Social Processes	11.4a	10.7ab	9.29b	.04
Humans (e.g. child, people, selves)	0.95ab	0.52a	1.12b	.05
Family (mother, father)	0.46ab	0.52a	0.25b	.08
Content				
Death (e.g. dead, killing, murder)	0.55	0.47	0.64	
Achievement	0.94	0.89	0.81	
Money (e.g. buy, economy, wealth)	0.34	0.38	0.58	
Religion (e.g. faith, Jew, sacred)	2.41	1.84	1.89	

Note. Numbers are mean percentages of total words per text file. Statistical tests are between Bin Ladin, Zawahiri, and Controls. Documents whose source indicates "Both" (n=3) or "Unknown" (n=2) were excluded due to their small sample sizes.

Advantage: Multi-lingual



APPENDIX B

DICTIONARY OF THE COMPUTER-BASED CONTENT ANALYSIS

	NL	UK	GE	IT
Core	elit*	elit*	elit*	elit*
	consensus*	consensus*	konsens*	consens*
	ondemocratisch*	undemocratic*	undemokratisch*	antidemocratic*
	ondemokratisch*			
	referend*	referend*	referend*	referend*
	corrupt*	corrupt*	korrupt*	corrot*
	propagand*	propagand*	propagand*	propagand*
	politici*	politici*	politiker*	politici*
	bedrog	*deceit*	täusch*	ingann*
	bedrieg	*deceiv*	betrüg*	
			betrug*	
	verraa	*betray*	*verrat*	tradi*
	verrad			
	schaam*	shame*	scham*	vergogn*
			schäm*	
	schand*	scandal*	skandal*	scandal*
	waarheid*	truth*	wahrheit*	verità
	oneerlijk*	dishonest*	unfair*	disonest*
			unehrlich*	
Context	establishm*	establishm*	establishm*	partitocrazia
	heersend*	ruling*	*herrsch*	
	capitul*			
	kapitul*			
	kaste*			
	leugen*		lüge*	menzogn*
	lieg*			mentir*

(from Rooduijn and Pauwels 2011)

Disdvantage: Highly specific to context

- Example: Loughran and McDonald used the Harvard-IV-4 TagNeg (H4N) file to classify sentiment for a corpus of 50,115 firm-year 10-K filings from 1994–2008
- found that almost three-fourths of the "negative" words of H4N were typically not negative in a financial context
  e.g. *mine* or *cancer*, or *tax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice*
- Problem: polysemes – words that have multiple meanings
- Another problem: dictionary lacked important negative financial words, such as *felony*, *litigation*, *restated*, *misstatement*, and *unanticipated*

Supervised Learning

Supervised Learning

- Dictionary methods require us to pre-define lists of words corresponding to classes of interest in advance
- Supervised learning allow us to assign documents to classes based on a training set of documents
- Advantages:
  - In contrast to dictionary methods, supervised learning is domain specific
  - Straightforward to validate, with familiar statistics for model performance

Supervised Learning

There are many, many models for supervised text learning. But all follow a few basic steps:

1. Define a training set of manually labelled documents, with each document assigned to a class
2. Apply a supervised learning algorithm in order to learn the relationship between features and classes in the training set
3. Validate the results using the test set (accuracy, precision, percent correctly predicted, etc)
4. Classify remaining documents in the full corpus

(If you attended day 4 of this course, this should sound familiar.)

Supervised learning example

Recall the Naive Bayes classification model from Thursday:

- Assume that features (words) are independent
- The probability of a document, d, being assigned to a class, k:

$$P(Y = k|X) \propto P(k) \prod_{i=1}^{n} P(x_i|k) \tag{2}$$

- We then assign the document to kth class for which it has the highest posterior probability:

$$\hat{y} = \underset{k \in \{1,\ldots,k\}}{\operatorname{argmax}} P(k) \prod_{i=1}^{n} P(x_i|k) \tag{3}$$

- Note that this is certainly not a correct model of text! The conditional independence of features assumption implies that we are essentially treating documents as bags of words

Supervised learning example

We are going to use the Naive Bayes model to predict whether movie reviews are positive or negative.

- Corpus: 2000 movie reviews
- Training set: 1000 reviews
- Test set: 1000 reviews
- Each review is manually labelled as either “positive” or “negative”

Naive bayes and quanteda

```
> # Load libraries
> library(quanteda)
> library(quantedaData)
>
```

Naive bayes and quanteda

```
> # Load libraries
> library(quanteda)
> library(quantedaData)
>
```

```
> # Convert corpus into document-feature matrix
> movie_dfm <- dfm(data_corpus_movies)
> dim(movie_dfm)
```

```
[1]  2000 48127
```

Naive bayes and quanteda

```
> # Load libraries
> library(quanteda)
> library(quantedaData)
>
```

```
> # Convert corpus into document-feature matrix
> movie_dfm <- dfm(data_corpus_movies)
> dim(movie_dfm)
```

```
[1]  2000 48127
```

That is a lot of features! Let's get rid of some of them:

```
> # Remove some features
> movie_dfm <- dfm_remove(movie_dfm, pattern = stopwords("english"))
> movie_dfm <- dfm_trim(movie_dfm, min_count = 10)
> movie_dfm <- dfm_trim(movie_dfm, min_docfreq = 0.02)
> dim(movie_dfm)
```

```
[1] 2000 2490
```

Naive bayes and quanteda

Let's subset the dfm into two, one for the training set, and one for the test set:

```
> # Subset the matrix into test and training sets
>
> test_set_vector <- docvars(data_corpus_movies)$test
> movie_dfm_test <- dfm_subset(movie_dfm,
+                               subset = test_set_vector)
> movie_dfm_train <- dfm_subset(movie_dfm,
+                                subset = !test_set_vector)
```

Now we can estimate the naive bayes model:

```
> # Estimate the naive bayes model on the training model
> nb_out <- textmodel_NB(movie_dfm_train,
+                        docvars(movie_dfm_train)$Sentiment)
```

Naive bayes and quanteda

Finally, we can predict the classes for the test set:

```
> test_set_predictions <- predict(nb_out, movie_dfm_test)$nb.predicted
> true_classifications <- docvars(movie_dfm_test)$Sentiment
> table(test_set_predictions, true_classifications)

                    true_classifications
test_set_predictions neg pos
                 neg 417  95
                 pos  96 392

>
```

Feature selection in Naive Bayes

We chose some pretty arbitrary rules for reducing the size of the DFM. But these choices can have important consequences.

1. Raw word counts (48,127 features)
2. Remove stop words (47,581 features)
3. Stem words (31,482 features)
4. Trim words (remove words that appear in less that 1% of documents, 556 features)
5. Sentiment only words (only words that appear in a dictionary of “positive” and “negative” words, 314 features)

(Note that these are still pretty arbitrary: why not 1 and 3? Or 2 and 4? etc)

Feature selection in Naive Bayes

Table: Accuracy for different feature selection strategies

DFM	Accuracy
Trimmed	0.787
Sentiment Only	0.787
Stemmed	0.969
Raw	0.981
Stop words	0.984

Feature selection in Naive Bayes

Table: Accuracy for different feature selection strategies

DFM	Accuracy
Trimmed	0.787
Sentiment Only	0.787
Stemmed	0.969
Raw	0.981
Stop words	0.984

It is a little weird that the ‘Sentiment Only’ model performs so poorly. Why is this?

Validation is important

It seems that our classifier is finding that particular films are good or bad, rather than finding something more general about the language used in movie reviews.

Table: Probability of class given word

	neg	pos
mulan	0.020	0.980
shrek	0.023	0.977
lebowski	0.032	0.968
leila	0.037	0.963
mallory	0.044	0.956
poker	0.045	0.955

This may have implications for scaling to other corpora of movie reviews!

Conclusion

- QTA allows us to draw inferences from very large collections of text without (too much) human interpretation
- All quantitative models of text are wrong, but some are useful
- Simple quantitative metrics of text can be very revealing
- Supervised text models, such as Naive Bayes, are easy to apply and can be very helpful in dealing with huge corpora
- `quanteda` is awesome
- Tomorrow: Unsupervised text models: topic models